Creating Knowledgebases to Text-Mine PUBMED Articles Using Clustering Techniques

Chiquito J Crasto, Ph.D.[1,2], Thomas M. Morse, Ph.D.[2], Michele Migliore, Ph.D.[2,7], Prakash Nadkarni, M.D.[1], Michael Hines, Ph.D.[2], Douglas E. Brash, Ph.D.[5,6], Perry L. Miller, M.D., Ph.D.[1,3,4] and Gordon M. Shepherd, M.D., D. Phil.[2]

[1]Center for Medical Informatics, [2]Department of Neurobiology, [3]Department of Anesthesiology, [4]Department of Molecular, Cellular, and Developmental Biology, [5]Department of Therapeutic Radiology, [6]Department of Genetics, Yale University, New Haven, Connecticut, USA. [7]Institute of Biophysics, National Research Council, Palermo, Italy.

Abstract

Knowledgebase-mediated text-mining approaches work best when processing the natural language of domain-specific text. To enhance the utility of our successfully tested program-NeuroText, and to extend its methodologies to other domains, we have designed clustering algorithms, which is the principal step in automatically creating a knowledgebase. Our algorithms are designed to improve the quality of clustering by parsing the test corpus to include semantic and syntactic parsing.

With a plethora of online resources becoming available, it is incumbent upon database developers and curators to be able to disseminate the rapidly accumulating data to researchers and users.

We recently developed NeuroText (http://senselab.med.yale.edu/textmine/neurotext.pl), a text mining program to automatically populate databases (http://senselab.med.yale.edu/senselab) in SenseLab (Crasto et al., 2003). NeuroText uses variable-scoring of identified database keywords (meta-data) to mark articles relevant for deposition into the databases. The scoring is based on contextual, semantic and syntactic constraints. These constraint-concepts are stored in a knowledgebase. NeuroText also incorporates two learning steps (supervised and unsupervised) to continually update the knowledgebase and enhance the text-mining efficacy. Approximately 3000 abstracts from the Journal of Neuroscience were scanned with an accuracy of 85%. NeuroText has been extended to PUBMED (http://chutney.med.yale.edu/textmine/Cerebellum_Purkinje.pl) neuroscience abstracts.

One of the drawbacks of a knowledgebase-mediated approach to text-mining is the time and expert-resource costs in creating a comprehensive knowledgebase. We developed algorithms that use traditional clustering concepts (Salton & Wong, 1978) to scan and cluster domain-specific PUMED abstracts. In addition to "bag of words" counting, our algorithms weight words based on semantic relationships identified within the Universal Networking Language paradigm (http://www.unl.ias.unu.edu/).

Our clustering techniques involve the following steps: Preprocessing the text to separate, weight and extract domain specific words from noise and stop words; combining a "vector space model" into a sparse matrix; and using the k-means algorithm to cluster the test set of PUBMED abstracts. (Dhillon, Fan, & Guan, 2001)

NeuroText was designed to be extensible to other domains without significant algorithmic modifications. This presentation will demonstrate the extension of NeuroText to a neuroscience-*related* domain- theoretical neuronal models and a non-related domain-p53 (tumor-suppressor gene) where the knowledgebase was created by clustering PUBMED abstracts. The expert/curator is the final arbiter for NeuroText results because of the necessity of depositing information into databases with 100% accuracy. The efficacy of our clustering methods and the results of the identifying relevant articles will be discussed in greater details.

References

1. Crasto, C. J., Marenco, L. N., Migliore, M., Mao, B., Nadkarni, P. M., Miller, P. L., & Shepherd, G. M. (2003). *Text Mining Neuroscience Journal Articles to Populate Neuroscience Databases*. Neuroinformatics. In Press.
2. Dhillon, I. S., Fan, J., & Guan, Y. (2001). *Efficient Clustering of Very Large Document Collections.* In V. Kumar & C. Kamath & R. Grossman (Eds.), *Data mining for scientific and engineering applications.* Dordrecht ; Boston, Mass.: Kluwer Academic.
3. Salton, G., & Wong, A. (1978). *Generation and Search of Clustered Files*. ACM Transactions on Database Systems. 2(4), 321-346.